

Chemometrics and model fitting in analytical chemistry

By

Matthew James Foot

This thesis is submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Department of Chemistry, Materials and Forensic Science
University of Technology, Sydney


2005

Certificate of Authorship / Originality

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all the information sources and literature used are indicated in the thesis.

Signature of Candidate



Acknowledgements

I would like to thank my principal supervisor Dr. Philip Maynard who stepped in as my supervisor two and a bit years into my project.

I would also like to thank my co-supervisor Associate Professor Les Kirkup. Les was always there providing help and support, and constantly challenging me to try and be a better scientist. Les's knowledge of experimental design and data analysis, as well as all the little tricks with Excel has been very useful throughout my time as a postgraduate student.

I would also like to thank my former principal and now unofficial supervisor Dr. Mary Mulholland. It was Mary who started me off down the path of research and helped steer me through the early part of this project. I am incredibly thankful that Mary made me write papers early on in my thesis especially at the end of my first year. Mary has continued to be very helpful even after leaving UTS and I am grateful that she gave up her time for me.

I greatly appreciate the help provided by Dr Peter Ghosh, who kindly gave me some samples used in my project and gave me some direction. I would also like to thank Susan Shimmon who showed me the lab at INR. I would also like to thank Ronald Shimmon who acted as a go-between for Dr Ghosh and me as well as being my friend and helping me with various questions throughout my time here.

I am constantly grateful for the friendship and support of Anthea Lloyd-Jones. Anthea has been a great support throughout my time as a research student, helping keep me informed about what happening and being there to bounce ideas off when needed. Anthea has also been an ear to listen to my woes for the last three and one half years, and for this I thank her greatly.

My fellow postgraduates are a bunch of champions and champion Uno players. Since moving into the office with them my life has been much more fun and much less stressful here at UTS. The Uno world championship gives me a bright spot to the day even when all things are going bad. So thanks heaps to Alison, Katherine, Sonia, Lisa, Bec, Tristan, Mark and Garry. I would also like to thank Sarka Prochazka, who helped me greatly when I started on this PhD road.

I would also like to thank Louise for being a great support since we met. Thanks for letting me stay at your place, and thanks for getting me out and doing things these last three years. Love you lots.

Without the help and support of my mum and dad, I would have never made it through these last three years. Indeed I thank them for getting me through university from when I started. I'd also like to say thanks to my brother; he didn't do all that much but he's my brother and he would feel left out if I didn't mention him.

List of publications

Some of the work presented in this thesis has been published in the following refereed journal articles

1. Foot, M., M. Mulholland, and L. Kirkup, *Classification of the biopolymer sodium pentosan polysulfate by infrared spectroscopy*. *Chromatographia*, 2003. **58**(1/2): p. 343-348.
2. Kirkup, L., M. Foot, and M. Mulholland, *Comparison of equations describing band broadening in high-performance liquid chromatography*. *Journal of Chromatography A*, 2004. **1030**(1-2): p. 25-31.
3. Foot, M. and M. Mulholland, *Classification of chondroitin sulfate A, chondroitin sulfate C, glucosamine hydrochloride and glucosamine 6 sulfate using chemometric techniques*. *Journal of Pharmaceutical and Biomedical Analysis*, 2005. **38**(3): p. 387-407.

Abstract

The aims of this project are to investigate the ability of advanced mathematical techniques and their contribution to the analysis of complex situations in analytical chemistry. The project falls into two areas. The first is the use of chemometrics to classify glycosaminoglycans (GAGs) such as chondroitin sulfates and glucosamines, substances that are being investigated for their potential use in the treatment of arthritis symptoms. This work is then expanded to the classification of novel anti-arthritis agents from different manufactures. The second part of this project looked at fitting different chromatographic band broadening models to real data. This work attempted to provide greater understanding of the processes involved in band broadening.

This classification work used different infrared spectroscopy techniques to analyse the molecules to which the classification systems were applied. Fourier Transform Infrared spectroscopy (FTIR), diffuse reflectance spectroscopy (DRIFTS) and Attenuated Total Reflectance spectroscopy (ATR) were evaluated for the classification of the chondroitin sulfates and glucosamines. The use of different spectral regions and derivative spectra were evaluated for their effect on the classification of the samples. It was found that FTIR coupled with derivative spectrums below 2000 cm^{-1} provided the best classification of these molecules.

The classification of sodium pentosan polysulfate was then considered using methods developed in the classification of chondroitin sulfates and glucosamines. Samples of the same material made by different manufacturers were provided to see if classification methods could distinguish them. Transmission spectroscopy coupled with the chemometric methods similar to those used for the classification of chondroitin

sulfate and glucosamines, were able to discriminate the samples by manufacturer and partially discriminate the samples by batch.

The second part of this thesis looks at fitting models to data as opposed to building classification models from data. This section looks at band broadening models in liquid chromatography followed by gas chromatography and finally looking at model fitting generally with some theoretical data. In this section a new model for band broadening in liquid chromatography is proposed. This model was found to be better able to predict band broadening behaviour in liquid chromatography at high flow rates. This model was then applied to gas chromatography; however it was found that previously published models best fit the data. The theoretical analysis highlighted the need for high quality data in order to draw conclusions about the model. It was also found that post column processes had the greatest effect on the band broadening.

Overall it was found that these advanced analytical techniques would be able to significantly improve the analysis of Glycosaminoglycan type compounds and provide further understanding of the band broadening process in liquid chromatography.

Table of contents

Acknowledgements	i
List of publications.....	iii
Abstract	iv
Table of contents	vi
List of figures	ix
List of tables.....	xiii
Chapter 1: Modelling and model fitting in analytical chemistry.....	1
1 Modelling and model fitting in chemistry.....	2
1.1 Multivariate analysis	2
1.2 Chemometrics	4
1.3 Computers and software.....	5
1.4 Chemometric Techniques.....	5
1.4.1 Exploratory data analysis, data reduction and principal components analysis. 5	
1.4.1.1 Principal components analysis	6
1.4.1.2 Hierarchical cluster analysis (HCA)	11
1.4.2 Pattern Recognition.....	12
1.4.2.1 Linear Discriminant Analysis	14
1.4.2.2 Soft Independent Modelling of Class Analogies (SIMCA)	17
1.4.2.3 Validation of pattern recognition models.....	19
1.5 Non-linear regression and model selection	19
1.5.1 Regression analysis	19
1.5.2 Model selection criteria.....	22
1.5.2.1 Akaike Information Criterion.....	23
1.5.2.2 Bayesian (Schwarz) Information Criterion	24
1.6 References	25
Chapter 2 : Glycosaminoglycans use and analysis.....	35
2 Glycosaminoglycans: use and analysis.	36
2.1 Current Treatments.....	37
2.1.1 Pharmaceuticals.....	37
2.1.1.1 Corticosteroids	37
2.1.1.2 Non steroidal anti-inflammatory drugs (NSAIDs).....	38
2.1.2 Natural products and remedies	38
2.1.3 Surgery	40
2.2 Glycosaminoglycans	40
2.2.1 Analysis of Glycosaminoglycans.....	42
2.2.1.1 Spectroscopy	43
2.2.1.2 Separations	48
2.2.1.3 Other techniques.....	54
2.3 References	55
Chapter 3: Classification of chondroitin sulfate A, chondroitin sulfate C, glucosamine hydrochloride and glucosamine 6 sulfate, using chemometrics and different infrared techniques	67

3	Classification of chondroitin sulfate A, chondroitin sulfate C, glucosamine hydrochloride and glucosamine 6 sulfate, using chemometrics with different infrared techniques.....	68
3.1	Introduction	68
3.2	Experimental	71
3.2.1	Chemicals and reagents.....	71
3.2.2	Transmission spectroscopy	71
3.2.3	DRIFTS	72
3.2.4	ATR.....	72
3.2.5	Data analysis	72
3.3	Results and Discussion.....	73
3.3.1	Spectroscopy	73
3.3.1.1	Chondroitin sulfate A and chondroitin sulfate C	75
3.3.1.2	Glucosamine hydrochloride and glucosamine 6 sulfate.....	75
3.3.2	Chemometrics	76
3.3.2.1	Transmission spectroscopy	76
3.3.2.2	DRIFTS	87
3.3.2.3	ATR.....	96
3.4	Conclusions	102
3.5	References	105
Chapter 4 : Classification of the biopolymer sodium pentosan polysulfate by Fourier transform infrared spectroscopy		110
4	Classification of the biopolymer sodium pentosan polysulfate by Fourier transform infrared spectroscopy (FTIR).....	111
4.1	Introduction	111
4.2	Experimental	116
4.2.1	Chemicals and reagents.....	116
4.2.2	Instrumentation	116
4.2.3	Data Analysis	116
4.3	Results and Discussion.....	117
4.3.1	Spectroscopy	117
4.3.2	Principal components analysis	120
4.3.3	Linear discriminant analysis	124
4.3.4	SIMCA	126
4.4	Conclusions	128
4.5	References	130
Chapter 5 : Fitting band broadening equation in chromatography		134
5	Fitting band broadening equations in chromatography.....	135
5.1	Analytical Chromatography	135
5.1.1	Columns	136
5.1.2	Detectors	136
5.2	Band broadening	138
5.2.1	Kinetic rate theory of chromatography	140
5.2.2	More than just van Deemter.....	141
5.3	Why model band broadening equations?.....	144
5.4	Experimental	145
5.4.1	Chemicals and reagents.....	145

5.4.2	HPLC analysis.....	146
5.4.3	GC analysis	146
5.4.4	Theoretical Analysis.....	147
5.4.5	Data Analysis	147
5.5	Results and Discussion.....	148
5.5.1	HPLC	148
5.5.2	GC	162
5.5.3	Theoretical Data	171
5.6	Conclusions	174
5.7	References	177
Chapter 6 : Conclusions		181
6	Conclusions and recommendations.....	182
6.1	Classification systems	182
6.2	Model fitting	183
6.3	Further Work	184

List of figures

Figure 1.1: A score plot of a FTIR data set showing the first two principal components plotted orthogonally to each other. The different coloured points represent different samples. This data has been mean centered.	8
Figure 1.2: A loadings plot of a principal component extracted from mid infrared FTIR data. The loading is plotted against the variables of that data set.	10
Figure 1.3: A dendrogram.	12
Figure 2.1: Common monosaccharide components of Glycosaminoglycans	41
Figure 2.2: Chondroitin sulfate A with the N-acetylgalactosamine ring sulfated in the 4 position joined to D-glucuronic acid.	42
Figure 2.3 Aggrecan: An example of a cartilage proteoglycan which contains both chondroitin sulfate and keratin sulfate.	42
Figure 3.1: Chondroitin sulfate A	69
Figure 3.2: Chondroitin sulfate C.	69
Figure 3.3: Glucosamine hydrochloride (left) and glucosamine 6 sulfate (right).	70
Figure 3.4: The transmission spectra of glucosamine 6 sulfate (top left), glucosamine hydrochloride (top right), chondroitin sulfate A (bottom left) and chondroitin sulfate C (bottom right)	74
Figure 3.5: First two principal components extracted from the full spectrum. The black squares represent chondroitin sulfate A, the yellow squares represent chondroitin sulfate C, The green circles represent glucosamine hydrochloride, and the blue circles represent glucosamine 6 sulfate.	77
Figure 3.6: Loadings plot from the first principal component extracted from the raw spectrum	77
Figure 3.7: Loadings plot from the second principal component extracted from the raw spectrum	78
Figure 3.8: Score plot of the first two principal components extracted from the spectrum below 2000 cm^{-1} The black squares represent chondroitin sulfate A, the yellow squares represent chondroitin sulfate C, The green circles represent glucosamine hydrochloride, and the blue circles represent glucosamine 6 sulfate.	79
Figure 3.9: First two principal components extracted from the full first difference spectrum The black squares represent chondroitin sulfate A, the yellow squares	

represent chondroitin sulfate C, The green circles represent glucosamine hydrochloride, and the blue circles represent glucosamine 6 sulfate. 79

Figure 3.10: First two principal components extracted from the first difference spectrum below 2000 cm^{-1} . The black squares represent chondroitin sulfate A, the yellow squares represent chondroitin sulfate C, The green circles represent glucosamine hydrochloride, and the blue circles represent glucosamine 6 sulfate. 80

Figure 3.11: Dendrogram created from the first three principal components of the full spectrum showing the clustering of the glucosamine hydrochloride (gluhcl), chondroitin sulfate A (CSA), chondroitin sulfate C (CSC) and the glucosamine 6 sulfate (G6S). ... 81

Figure 3.12: Dendrogram of the first three principal components above 2000 cm^{-1} showing the clustering of the glucosamine hydrochloride (gluhcl), chondroitin sulfate A (CSA), chondroitin sulfate C (CSC) and the glucosamine 6 sulfate (G6S). 82

Figure 3.13: First three principal components extracted from the raw DRIFTS spectra The black squares represent chondroitin sulfate A, the yellow squares represent chondroitin sulfate C, The green circles represent glucosamine hydrochloride, and the blue circles represent glucosamine 6 sulfate. 88

Figure 3.14: Loadings plot for principal component 2 90

Figure 3.15: Score plot of principal component 2 vs. principal component 3 extracted from the first difference DRIFTS spectra above 2000 cm^{-1} . The black squares represent chondroitin sulfate A, the yellow squares represent chondroitin sulfate C, The green circles represent glucosamine hydrochloride, and the blue circles represent glucosamine 6 sulfate. 90

Figure 3.16: Loadings plot of principal component 3 91

Figure 3.17: Dendrogram obtained using only principal component 2 and principal component 3 showing the clustering of the glucosamine hydrochloride (GH), chondroitin sulfate A (CA), chondroitin sulfate C (CC) and the glucosamine 6 sulfate (GS). 92

Figure 3.18: An ATR spectrum of chondroitin sulfate A 97

Figure 3.19: 3d score plot of ATR raw spectrum. The black squares represent chondroitin sulfate A, the yellow squares represent chondroitin sulfate C and the green circles represent glucosamine hydrochloride. 98

Figure 3.20: PC2 vs. PC3 score plot from ATR data. The black squares represent chondroitin sulfate A, the yellow squares represent chondroitin sulfate C and the green circles represent glucosamine hydrochloride. 99

Figure 3.21: Raw spectra dendrogram showing the clustering of the glucosamine hydrochloride (GH), chondroitin sulfate A (CA) and chondroitin sulfate C (CC). 100

Figure 3.22: First difference dendrogram glucosamine hydrochloride (GH), chondroitin sulfate A (CA) and chondroitin sulfate C (CC).	100
Figure 4.1: The repeating unit of NaPPS [2].....	111
Figure 4.2: A typical near infrared spectrum of NaPPS.....	118
Figure 4.3: A typical mid infrared spectrum of NaPPS	118
Figure 4.4: %RSD plots from the near infrared spectra. The lines for each sample indicate the intra-batch variation, while the combined RSD indicates the %RSD across all samples.	118
Figure 4.5: %RSD plots from the mid infrared spectrum. The lines for each sample indicate the intra-batch variation, while the combined RSD indicates the %RSD across all samples.	119
Figure 4.6: An overlay of typical MIR spectra obtained from the two manufacturers. Manufacturer A is in blue and manufacturer B is in red.....	120
Figure 4.7: A score plot of the first two principal components of the entire raw spectrum.	121
Figure 4.8: The loadings plot for the first principal component, obtained for the full spectrum.	121
Figure 4.9 Score plot of the first two principal components of the raw spectrum using only the region $<1800\text{ cm}^{-1}$	122
Figure 4.10: Score plot of the first two principal components of the first difference spectrum.	123
Figure 4.11: Score plot of the first two principal components of the first difference spectrum using only the region below 1800 cm^{-1}	123
Figure 5.1: Derivation of H from σ^2 and L from a chromatographic peak.	139
Figure 5.2: H vs. u data for propylparaben collected using HPLC.	149
Figure 5.3: Residuals for unweighted fit of equation (5.7) to data in Figure 5.2.....	149
Figure 5.4: Residuals for weighted fit of equation (5.7) to data in Figure 5.2.....	150
Figure 5.5: Propylparaben H vs. u data with equations (5.3) (5.5), (5.6) and (5.7) fitted.	151
Figure 5.6: Weighted residuals of equations (5.3), (5.5) and (5.7) plotted against linear velocity for propylparaben data.	151
Figure 5.7: Plot of HETP vs. linear velocity (0.1 spectra/s)	157

Figure 5.8: Plot of HETP vs. linear velocity (1 spectra/s)	159
Figure 5.9: Plot of HETP vs. linear velocity (10 spectra/s)	159
Figure 5.10: Standard deviation versus flow rate for 0.1 spectra/s sample rate and 10 spectra/s sample rate	160
Figure 5.11: 15 ms exposure time H versus u plot.....	161
Figure 5.12: 150 ms exposure time H versus u plot.....	161
Figure 5.13: Standard deviation vs. flow rate for both exposure times.	162
Figure 5.14: H versus u plot of toluene analysed by GC sampling rate 200 Hz.....	163
Figure 5.15: Unweighted (top) and weighted (bottom) residuals for toluene from fit of equation (5.5).	164
Figure 5.16: Histogram of “best fit” equations at 1% noise	172
Figure 5.17: Histogram of “best fit” at 3% noise.....	173
Figure 5.18 : Best fit determined by BIC at 6% noise	173
Figure 5.19: AIC at 9% noise.....	174

List of tables

Table 3-1: LDA on spectrum below 2000 cm ⁻¹	84
Table 3-2: A summary of LDA results for transmission spectroscopy.....	84
Table 3-3: Number of outliers, principal components and % variance explained for each spectral region using raw spectra.	85
Table 3-4: Class distances calculated for each spectral region using the raw spectra	86
Table 3-5: Number of outliers, principal components and % variance explained for each spectral region using the first difference spectra.....	87
Table 3-6: Class distances calculated from each spectral region of the first difference spectra.	87
Table 3-7: A summary of LDA classification rates using DRIFTS spectra.	93
Table 3-8: Number of outliers, principal components and % variance extracted for each group using the DRIFTS spectra.....	95
Table 3-9: Class distances calculated for each spectral region using the DRIFTS spectra	95
Table 3-10: Number of outliers, principal components and % variance extracted for each group using the first difference DRIFTS spectra.....	96
Table 3-11: Class distances calculated for each spectral region using the first difference DRIFTS spectra.....	96
Table 3-12: Cross validated LDA on the raw spectrum.....	101
Table 3-13: Cross validated LDA on the first difference spectrum	101
Table 3-14: Number of principal components and the % variance explained for the cross validated models.....	102
Table 3-15: Class distances calculated using the cross validated models.....	102
Table 4-1: List of samples.....	116
Table 4-2: Overall classification rate for NaPPS samples using LDA on the first three principal components. Samples were classified to the different manufacturers.	124
Table 4-3: Classification of batches using LDA on raw spectrum below 1800cm ⁻¹	126
Table 4-4: Classification of batches using LDA on the first difference spectra below 1800cm ⁻¹	126

Table 4-5: Class distances calculated using the full spectral range and the spectrum below 1800 cm ⁻¹ for the raw spectra.	127
Table 4-6: Class distances calculated using the full spectral range and the spectrum below 1800 cm ⁻¹ for the first difference spectra.	128
Table 5-1: Parameter estimates from the weighted and unweighted fits of equations (5.3) and (5.5)-(5.7).....	152
Table 5-2: Fitting statistics from methylparaben and propylparaben.	153
Table 5-3: Predictive ability of equations (5.3), (5.5), (5.6) and (5.7).....	155
Table 5-4: Different experimental conditions for investigation of source of scatter and curvature.....	156
Table 5-5: Model fitting data	158
Table 5-6: Parameter estimates for GC data at 20 Hz sampling rate	165
Table 5-7: Parameter estimates for GC data at 100 Hz.....	166
Table 5-8: GC parameter estimates for GC data at 200 Hz	167
Table 5-9: Fitting statistics for GC data.....	168
Table 5-10: predictive ability of the equations	170